LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 12 (2018), pp. 204–234
http://nflrc.hawaii.edu/ldc
http://hdl.handle.net/10125/24768
Revised Version Received: 17 Jan 2018

A Guide to the Syuba (Kagate) Language Documentation Corpus

Lauren Gawne
SOAS University of London
La Trobe University

This article provides an overview of the collection "Kagate (Syuba)", archived with both the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) and the Endangered Language Archive (ELAR). It provides an overview of the materials that have been archived, as well as details of the workflow, conventions used, and structure of the collection. It also provides context for the content of the collection, including an overview of the language context, and some of the motivations behind the documentation project. This article thus provides an entry point to the collection. The future plans for the collection – from the perspectives of both the researcher and Syuba speakers – are also outlined, but with the overwhelming majority of items in the collection available to others, it is hoped that this article will encourage use of the materials by other researchers.

1. Introduction Language documentation involves the development of corpora of materials from which descriptions of grammar and language use can be developed, alongside other uses of the materials by both speakers of the language and researchers. Himmelmann (1998) argues that language documentation and description are two distinct, but interrelated, activities. In reality, the majority of basic linguistic description based on primary data is undertaken by the same person who collected the data. Very little of this descriptive work makes clear the nature of the data on which it is built; in a survey of 50 published grammars and 50 PhD dissertations, Gawne et al. (2017) found that 68% gave no indication of whether the data had been archived, and many failed to make clear other basic features of methodology. A clear description of the nature of the data used in an analysis, and the ability for the reader to access the data, allow for greater reproducibility in language documentation, as a reader can attempt to reproduce the claims made by the original author, or ask their own questions of the data. It also reminds the reader of the centrality of the primary data to the research endeavor, as Himmelmann (1998:164) has noted, as have Thieberger (2009), Thieberger & Berez (2012), and Woodbury (2003), among others. This article describes the Syuba language documentation project, and the archived collection of materials from the project. This fits with current discussions within the field of language documentation about the nature of the data archives that we are building, including Austin's (2013) call for broader metadata regarding the language documentation process, and Woodbury's (2014) call to make archived collections

E-ISSN 1934-5275

more accessible to potential users. I am not, by any means, the first to present this kind of information in peer-reviewed article format. Other scholars have used journal articles to provide supporting information about digital collections of linguistic material, at various levels of archival structure. Bouda & Helmbrecht (2012) provides an introduction to the 50+ language collections in the DoBeS archive, helping researchers to navigate the DoBeS site and using the materials found there. Barth & Evans (2015) provides an overview of The Social Cognition Parallax Interview Corpus (SCOPIC), a cross-linguistic collection archived with Language Archive Cologne[1] that replicates an interactive task to generate examples of features of grammar that relate to social interaction. Individual researchers will subsequently provide analyses from the collection as part of a rolling edited volume. For articles describing single-language documentation corpora, see Schembri et al. (2013) for British Sign Language and Salffner (2015) for Ikaan (Niger-Congo, Nigeria).

The Syuba collection has been deposited with both PARADISEC[2] and ELAR.[3] In ELAR the collection is called "Kagate (Syuba), an endangered Tibeto-Burman language of Nepal", which reflects the name of the ELDP project that the ELAR collection was established to support. In §2 I provide the background to the community and language (§2.1), the documentation project (§2.2), the data conventions used in the project (§2.3), and details of related digital collections (§2.4). In §3 I provide an overview of the online collection, before looking at the different types of material that can be found within it, including narrative and conversation (§3.1), song (§3.2), materials with a linguistic focus (§3.3), materials that document the project (§3.4), and other items (§3.5). Finally, I discuss future plans for the collection, including current research (§4.1), potential future uses of the corpus (§4.2), and ways to use and cite the materials in the collection (§4.3). This collection also provides the only documentation made to date of the Ilam variety of Yolmo, and I discuss these materials specifically in (§3.6). In providing this outline I hope to encourage other researchers to use the materials created, and I also hope to encourage other researchers with language documentation materials to provide contextualizing metadata in an article such as this. I return to these thoughts in the conclusion (§5).

2. Background

2.1 The Syuba people and their language Syuba is a Tibeto-Burman language spoken in the Ramechhap district, in the central eastern hills of Nepal. The language is mutually intelligible with at least some of the Yolmo[4] dialects spoken in other areas of Nepal, and is part of the Southern group of the Tibetic languages (Tournadre 2014), along with other languages on both the Nepal and Tibetan sides of the Himalaya including Nubri, Tsum, and Kyirong.

[1]https://lac.uni-koeln.de/en/

[2]http://catalog.paradisec.org.au/collections/SUY1

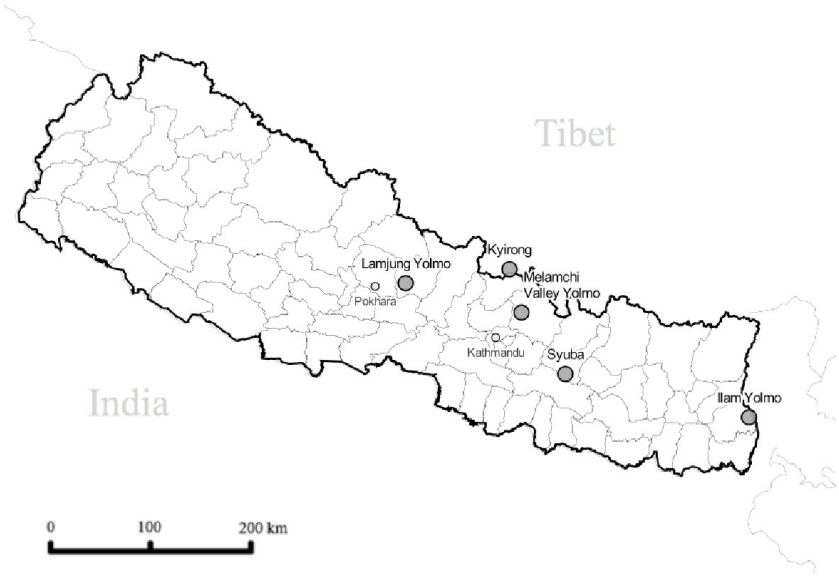
[3]http://elar.soas.ac.uk/deposit/0388

[4]Yolmo is also known as Helambu Sherpa, and is also spelled Yohlmo, Hyolmo or Yholmo, with the <h> included to represent the low tone on the word.

Syuba is the name used for both the language and the community, with some people also using it as part of their name, and is an endonym. Government records, and the existing linguistic literature on the language, refer to the language as *Kagate*, which is an exonym. Like *Syuba*, it means 'papermaker', in reference to the historical occupation of Syuba community members. The name Kagate is still recognized in the community, but there is a growing preference for using *Syuba*, as part of a growing interest in the language amongst speakers (see Gawne 2016c).

Although Syuba is mutually intelligible with other Yolmo languages, speakers see themselves as having a distinct ethnic identity and a separate language. The largest Yolmo population is spread through the Melamchi and Helambu Valleys north of Kathmandu (see Figure 1). From this main population three large groups, and likely many others, have migrated and settled in other areas of Nepal; one group are the Syuba, who settled in Ramechhap. The other two groups settled in Lamjung and Ilam, and refer to their own languages as Lamjung Yolmo and Ilam Yolmo, respectively. As in the other Yolmo groups, the majority of Syuba speakers practice Buddhism of the Tibetan *Nyingma* tradition; however, since the 1970s, there are also a small number of Christians. Figure 1 is a map of Nepal that shows the Syuba in Ramechhap, as well as other populations of Yolmo speakers, and Kyirong, which is a closely related language to the Yolmo languages (c.f. Hedlin 2011).

Figure 1. A map of Syuba and related languages



Oral histories in Ramechhap state that the Syuba have lived in the area for one to two centuries. There are currently around 1,500 speakers of Syuba in Ramechhap (Mitchell & Eichentopf 2013:3). The language has been classified as 6a on the Expanded Graded Intergenerational Disruption Scale (EGIDS) (Lewis & Simons 2010) by Mitchell & Eichentopf in a 2013 survey. This classification is used for languages with strong oral transmission across all generations in the home, but no sustained

literacy. The authors also reported a positive attitude by speakers towards their own language. I would agree that this assessment realistically reflects the state of Syuba use today. For more detail on the history and social context of Syuba and the people who speak the language, see Gawne (2017a).

2.1.1 Linguistic and geographic setting Syuba villages are located in multiple Village District Committee (VDC) zones (the smallest local administrative unit) in the Ramechhap district of Nepal. The villages are located in Nepal's "hills".[5] They are often the highest settlements on the slopes. These villages in Ramechhap are exclusively populated by Syuba speakers. There are around eight villages that are contiguously located in the Duragaun VDC, and a further half a dozen in the Namari and Bhuji VDCs.

Mitchell & Eichentopf (2013:3) report that the individuals they interviewed indicate no difference in accent across Syuba, although their survey did not include speakers from all areas. My work has mostly focused on the villages of the Duragaun VDC, and to date I have observed no major variation in language use between people from the different villages.

Other local communities include Sunwar[6] and Tamang;[7] these are both Tibeto-Burman languages, but from different branches. Sunwar is Kiranti and Tamang is Tamangic, and are mutually unintelligible with Syuba. There are also Nepali-speaking communities of Chhetri and Brahmin Hindus. Syuba speakers are also bilingual in Nepali, which is the language of education and communication with other groups. Individuals may also have some degree of competency in other languages of the region, or English.

2.1.2 Existing work on Syuba There are three main milestones in previous work on Syuba. The first is Grierson's (1909/1966) "Linguistic Survey of India", under the name "Kagate", because speakers of Syuba were living and working in Darjeeling at the time of the survey. The second is documentation work carried out by SIL from 1972 to 1976. Monika Höhlig undertook preliminary documentation of Syuba, development of an orthography, and translation of some Bible passages, along with preliminary linguistic analysis (Höhlig & Hari 1976; Höhlig 1978). Höhlig and Hari also produced an unpublished typewritten wordlist, which I scanned, digitized, and turned into a Toolbox wordlist (see Gawne 2014 for more details). Monika Höhlig also made recordings on reel-to-reel tapes from 1972–1976, which I have assisted her in digitizing and archiving as an open access collection at PARADISEC.[8]

Most recently, Syuba speakers have reconnected with SIL to recommence literacy development work. In February 2014, a 10-day orthography development workshop was hosted by the Mother Tongue Centre Nepal (MTCN),[9] and a Devanagari-based

[5]At elevations over 1800m, these "hills" are taller than what many countries would refer to mountains.
[6]ISO 639-3 suz, Glottocode sunw1242
[7]ISO 639-3 taj, Glottocode east2347
[8]This collection will be online in September 2017: http://catalog.paradisec.org.au/collections/MH1.
[9]http://mtcn.org.np/index.php/our-projects

orthography was established. In December 2013 and January 2014 the MTCN facilitated a language documentation workshop, where over 12 hours of audio recordings in a range of genres were made with sixty speakers of the language. Some written transcription and Basic Oral Language Documentation (BOLD) was done at the MTCN. I am currently working to archive this collection with PARADISEC, in agreement with the MTCN.[10] This project has texts in similar genres to those in the collection described in this paper, but efforts were made to reduce duplication. The collection was also predominantly audio-only, with only two videos. SIL also conducted a dictionary workshop[11] and a Syuba-Nepali-English dictionary is now available online,[12] with a printed version made for Syuba speakers (SIL International & HIS Nepal 2016).

2.1.3 A very brief introduction to features of Syuba grammar The descriptive work on Syuba phonology and grammar is still limited to two publications from the 1970s (Höhlig & Hari 1976; Höhlig 1978) and some initial observations from the current author (Gawne 2013b). Familiarity with descriptive work on the closely related Yolmo varieties provides a good introduction to many of the features that are found in the Syuba language.

There are two tones (low and high) which occur on the initial syllable of lexical words. These appear to occur in restricted environments based on the value of the initial consonant in a word. Voiced stops and affricates occur with low tone, aspirated voiceless aspirated stops and affricates occur with high tone, and in all other environments both are possible. Syuba appears to be more conservative than other Yolmo varieties, retaining word final /k/ on many words where it has been lost in other varieties, e.g., *ʧùk* 'six' rather *ʧù* for 'six' in Lamjung and Melamchi Yolmo.

Word order is SOV; however, in speech, there may be many referents that are not overtly marked. Case is marked with nominal clitics, with ergative marking appearing to be pragmatically motivated. Verbs have lost all but vestigial elements of the stem alteration found in other Tibetan varieties. Negation is the only function that is prefixed to the verb. Other tense, aspect, mood, and evidential functions are marked with verb suffixing and auxiliary verbs. There is a set of evidential distinctions that include a sensory evidential and a reported speech evidential particle. There are also other attitudinal particles that are used to mark interactional discourse. As has been noted by Höhlig & Hari (1976:1), the language lacks a great deal of the honorific vocabulary found in Melamchi Yolmo for talking about Lamas and other important people. The base-20 counting system is still used by some speakers, but a base-10 system is also in use.

2.2 The Syuba language documentation project

2.2.1 Motivation and set-up While working on the documentation of Lamjung Yolmo for my PhD dissertation, I contacted Anna Maria Hari to ask if she was still

[10]This collection will be online in September 2017: http://catalog.paradisec.org.au/collections/MTC1.
[11]http://rapidwords.net/report/syuba
[12]http://syuba.webonary.org

in touch with any Kagate speakers, as I was interested in the similarity between the two languages and the commonalities in their history of migration. Hari put me in touch with Norpu Lama, who was a young man when he had worked with Hari and Monika Höhlig in the 1970s, and was now the minister of a Christian church in Kathmandu. Throughout my PhD I stayed in touch with Norpu and other Syuba speakers who were living in Kathmandu. They expressed interest in documenting their own language and making recordings of oral history, songs, and stories.

After completing my PhD, we worked together to develop a project to document Syuba. The overall aim was to develop a corpus that reflected the Syuba speakers' interests in a historical documentation of the language and its speakers, but that would also be of use to researchers interested in the linguistic features of the language. My research interests in Syuba were threefold at the time the project commenced: (1) to establish the linguistic and historic relationship between Syuba and other Yolmo varieties, (2) to understand the evidential system of Syuba and its use in interaction, expanding on my work on Lamjung Yolmo, and (3) to develop a corpus of narrative and interactional uses of co-speech gesture in the language. Hence, audio-video documentation of narrative and oral history was also of interest to me. Subsequent collaborations with colleagues led to two lexical-tone experiments being designed and conducted during the project, as well as a number of other small projects within the larger scope of the documentation, which are mentioned below.

This work has been done under the auspices of three separately funded projects, but together they make up the completed Syuba language documentation that is in the online collection. The first grant was from The Firebird Foundation. This funded basic recording equipment and one month of fieldwork time in Nepal. Three and a half hours of audio and video recordings were made. During this fieldtrip I also took mobile phones with the Aikuma mobile application, which is built around the Basic Oral Language Documentation (BOLD) method (Reiman 2010) to test for Steven Bird and his project team at The University of Melbourne (see Bird et al. 2014). During this trip we also received small grants from The Awesome Foundation (Ottawa chapter)[13] and Stack Exchange.[14] This allowed us to purchase a camera, computer, and audio recorder to do documentation work.

The second portion of project funding was through Nanyang Technological University (NTU), where I was a College of Humanities, Arts, and Social Sciences Postdoctoral Research Fellow from 2014–2015. Some of the funding was associated with my postdoctoral position, and other funding was drawn from a collaborative grant with Joan Kelly in the NTU School of Art, Design and Media (ADM).[15] This funding allowed us to record 3.5 hours of audio and video, and also work with students from NTU AMD to develop picture books using the Syuba narratives. The first, made by Jolene Tan, illustrates the story of the Jackal and Pheasant, in which they trick the humans out of their food. The second, made by Ng Xiao Yan, is a compendium of

[13]www.awesomefoundation.org/en/chapters/ottawa

[14]www.stackexchange.com

[15]'The development of Artistic and Participatory Means of Recording, Writing and Transmitting the Stories and Knowledge of Kagate, an Endangered Language of Nepal' (Tier 1 grant).

eight different stories and songs. The project also allowed Joan Kelly and Ng Xiao Yan to visit Nepal to conduct drawing workshops with Syuba-speaking children.

The third portion of funding was through an Endangered Language Documentation Programme Postdoctoral Fellowship that I held from 2015–2017. This funding allowed me to spend five months in Nepal, recording over 10 hours of audio-video, conducting two different experiments, transcribing texts, and training Syuba speakers to also transcribe and translate texts.

As this project was developed because of community interest in documenting their own language, community members participated at many levels of the project. Community members participated in the recordings, arranged and conducted interviews with other community members, and ensured that a representational range of songs and stories were recorded. One speaker (Ningmar Tamang) was trained in audio and video recording, and did much of the recording during the ELDP project. Ningmar, along with Sangbu Syuba, was taught to use ELAN and segmented, transcribed, and translated recordings on computers provided. Ningmar has subsequently trained other speakers in the use of ELAN. During the NTU-funded project, two workshops were held for children, one in Kathmandu and one in Ramechhap. Children drew illustrations of animals with labels added in Syuba, which were collated into books as literacy resources. At the end of the project, two computers and sets of headphones were left with the speakers to continue recordings. At various points a camera, a Zoom H1, and two mobile phones with good photo cameras were also left with speakers so that they could continue their own recordings.

2.2.2 Project team The following people were involved in the documentation at various stages, other than as participants in recordings:

- **Lauren Gawne** Linguist and principal investigator.
- **Norpu Lama** Syuba speaker. Along with his family helped me to meet other Syuba speakers and design the initial Firebird Foundation grant.
- **Sangbu Syuba** Syuba speaker, a younger brother of Norpu Lama. Arranged village travel and recording schedule. Transcribed and translated recordings.
- **Ningmar Tamang** Syuba speaker. Recording assistant, who undertook almost all of the audio and video recording work in 2016. Transcribed and translated recordings.
- **Joan Kelly** Senior Lecturer and artist at NTU. Led the student drawing workshops in Nepal in 2016. Led the student-created storybook project at NTU in Singapore.
- **Ng Xiao Yan** and **Jolene Tan** Students at NTU ADM who created illustrated books from Syuba materials. Xiao Yan illustrated a compendium of eight songs and stories, and participated in the drawing workshops in Nepal. Jolene illustrated the story of the Jackal and Pheasant.

- **Suzy Styles** Assistant professor at NTU, a psycholinguist who collaborated with Lauren Gawne on the tone cross-sensory perception experiment.
- **Amos Teo** A PhD candidate at the University of Oregon. Collaborated with Lauren Gawne on the tone listening experiment. Also collaborated on the earlier digitization of the Höhlig and Hari typewritten dictionary and preliminary recordings with Norpu Lama on a 2010 field visit (see Gawne 2014).

All Syuba participants and Lauren Gawne were conversant in Nepali and most work activities were undertaken in Nepali, with some Syuba also used. Other international researchers used English, with Lauren Gawne, Ningmar Tamang, and Sangbu Syuba interpreting when in Nepal.

2.2.3 Impact and contribution Impact and contribution of this documentation project and the collection is relevant to three distinct (though not mutually exclusive) categories of audience: Syuba community members, academic audiences, and other audiences.

People who participated in recordings were made aware that the recordings were intended for open access archiving. This was partly motivated by my interest as a researcher in ensuring that the examples of language use I discuss are accessible to an academic audience, but was largely motivated by the desire of many Syuba speakers I talked with that their language become more visible to others. In creating a corpus of audio and video of the language and making it open access through two different archives, as well as creating open access archives for the two earlier documentation projects, we have increased the visibility of Syuba. Copies of all recordings were left with the community, mostly on micro-SD cards so that they could be shared on mobile phones. Copies of the picture story books with text in Syuba were also left, along with books containing the animal illustrations done by children as part of the drawing workshops.

The academic impact of this research is only just beginning to be demonstrated. The video data from this project has formed the basis for some of the first gesture analysis to be undertaken in Nepal (Gawne 2016b; 2017b). Both phonetics experiments are in the analysis stage with planned publications, the first of which is Styles & Gawne (2017). Initial work is now being undertaken on the distribution of evidentiality and other grammatical features.

Throughout this project we have also aspired to engage a broader audience beyond the Syuba speech community and academics. The Syuba book project at NTU ADM involved creating versions of these stories in Syuba, as well as bilingual copies in Syuba and either Nepali or English. This means that audiences in Nepal and abroad can also engage with Syuba traditions, stories, and songs. The books will be set up as print-on-demand, with any profit from the printing going towards creating additional copies for Syuba children. The project also provided two students at NTU with experience working on a real-world project where the output can be used by others, and required them to learn to produce works for a social context very different from their own.

2.3 Data conventions in the project

2.3.1 Data recording formats Recordings for this project were made from 2009 to 2016.[16] Different recording equipment was available at different points in the project. For each recording the equipment is listed in the metadata of the archive.[17] Audio recordings for this project were made using a Zoom H4n (WAV, 44.1kHz, 16-bit) except for a small subset that were made to test the Aikuma mobile phone software, which were recorded as WAV on a HTC Desire 210 using the internal microphone. For the 2009–2013 recordings the Zoom's built-in microphone was used. For the 2014 recordings a Sennheiser omnidirectional lavaliere microphone was used. For the 2016 recordings an Audio-Technica AT8022 stereo microphone was used. All video was recorded on a Panasonic HC-V720 video camera with a Rode Pro shotgun microphone to provide a decent quality backup audio. Video recordings were made in AVCHD, which provided uncompressed MTS files for archiving. The majority of still images were taken using a Canon Ixus 100is. A number of photographs were taken by other people on their own cameras. Where this is the case, attribution is given in the image metadata. Handwritten notes were scanned as PDFs.

2.3.2 Metadata and file naming Metadata for all files is included with each individual item in the online collections, and is also collated in a CSV file (SUY1-metadata-01). File names are based on a standard structure of three parts. The first is the collection level ID created for PARADISEC (SUY1). The next section is the date the recording was made in ISO format; *SUY1-091125* items were recorded on the 25th of November 2009, *SUY1-160511* was recorded on the 11th of May 2016. The final element is the sequence number of a recording on any given day; *SUY1-140127-01* is the first recording made on the 27th of January 2014, and *SUY1-140127-14* is the fourteenth recording made on that (rather long) day. Where the item in the collection is not a recording, the tripartite naming structure still applies, but the middle section indicates the type of file, for example *SUY1-images-5261* is an image, with the sequence number 5261, and *SUY1-metadata-01* is the metadata sheet for the whole collection.

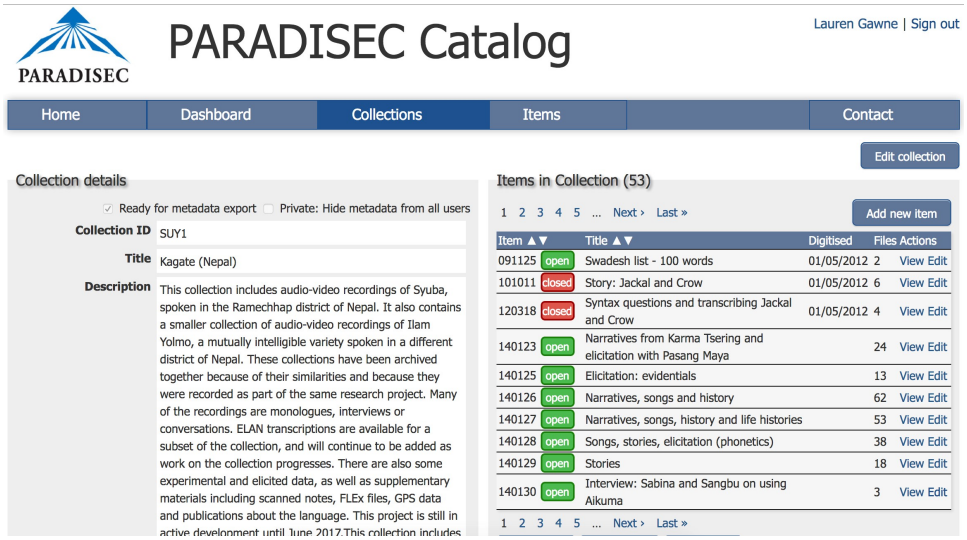
2.3.3 Annotations Written annotations have been made using ELAN[18] (Wittenburg et al. 2006) by myself, Ningmar Tamang, and Sangbu Syuba. Written translations and transcriptions have been produced for 84 recordings (totaling 550 minutes), and are still being developed for the rest of the corpus. Of those recordings that have been transcribed, 38 (223 minutes) have been interlinearized in FLEx[19] (SIL International 2017) and re-imported into ELAN. Recordings that have been interlinearized have been transcribed in a Latin-based orthography, the same as was used for Lamjung

[16] Any future recordings made will also be added to the archives.
[17] This information is only available at present through the PARADISEC archive.
[18] http://tla.mpi.nl/tools/tla-tools/elan/ Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands
[19] http://fieldworks.sil.org/flex

Yolmo (Gawne 2016a), and translated into English. The majority of the others have been transcribed by Ningmar using the Syuba Devanagari-based transcription system and translated into Nepali. Those transcribed by Sangbu Syuba use a Latin-based orthography of his own devising. References to annotations are based on their timecode in the recording. Only the EAF transcription files have been archived with PARADISEC. ELAR requests that the PFSX files that ELAN creates are also archived, and so these settings files are available through ELAR.

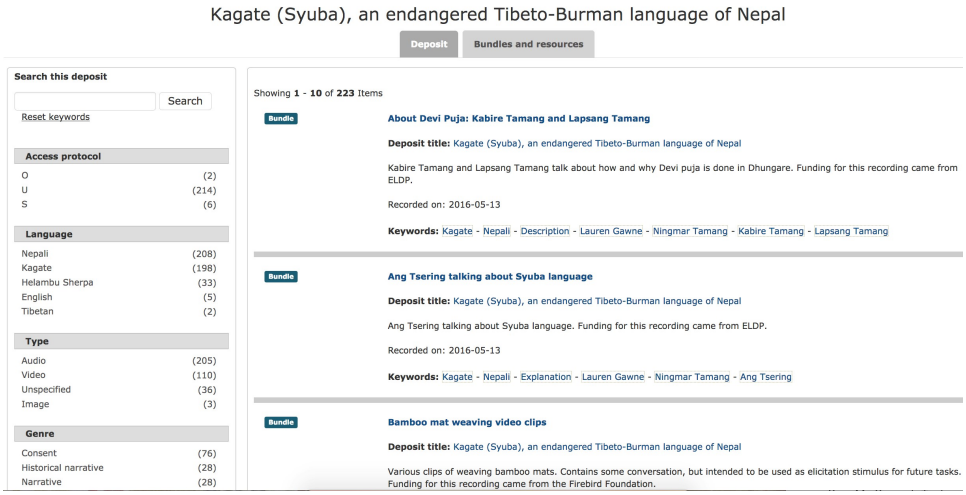
2.3.4 Archive structure The materials for this project were originally archived with PARADISEC. When the ELDP funding was awarded, it was decided to archive the entire collection with ELAR as well as PARADISEC, and not limit the ELAR collection to only a subset of the materials. This means that this collection offers an elegant illustration of the differences in how the two archives display the same data set. The way that the archives are structured has important implications for how the collection has been developed, and how it is displayed. PARADISEC creates "bundles" based on the middle part of the file name, that is, all files that begin with *SUY1-140127* are bundled together as a single item, and because the bundles are date-based in this collection, all recordings made on a single day are grouped together. The list of files is arranged by file name, so the oldest recordings are at the top of the collection, and the most recent are at the bottom. Figure 2 is a screenshot of the current view of the collection in PARADISEC.

Figure 2. The Syuba collection in PARADISEC



For ELAR, each individual recording event is contained in a single bundle of associated audio, video, and transcription files. This means that there are more "bundle"-level items in the collection. ELAR currently arranges records by the arbitrary bundle title, rather than the recording ID or filenames as per PARADISEC. This means that the files are displayed in this alphabetical order, rather than the chronological order at PARADISEC. Figure 3 is a screenshot of the initial records in the ELAR collection.

Figure 3. The Syuba collection in ELAR



A major limitation with ELAR's current structure is that it is not possible to search the collections by recording ID, only by title. This is a limitation because citation of the corpus in publications is done by the recording ID. The current ELAR workflow relies on The Language Archive (TLA)[20] and its Arbil metadata builder (Withers 2012),[21] and is much slower and more onerous for the researcher than PARADISEC's workflow, which is built on their own Nabu platform.[22] For these reasons, the PARADISEC collection should always be considered to be the most up to date, and the preferred link to cite when referring to the collection.

The majority of the collection is available openly to those who sign up for an account with the archive and agree to the conditions of access. In both archives there are a small number of items that are closed access. Where items are not open access I have made a note of this, and the reason why. These materials can still be made available to individual speakers of Syuba, or other researchers, where relevant.

[20]https://tla.mpi.nl

[21]http://tla.mpi.nl/tools/tla-tools/arbil Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands.

[22]https://researchdata.ands.org.au/nabu-catalogue/504365

2.4 Related collections There are a number of collections in PARADISEC that are of direct relevance to the Syuba collection. The collection described here is the most complete in terms of amount transcribed and available with video and audio, but the other collections include data that is not within the scope of this collection.

The first is an open access collection of historical materials from Syuba made in the 1970s by Monika Höhlig (MH1).[23] The digitized materials include recordings from three reel-to-reel tapes and six cassettes of Syuba from 1972–1976, as well as typed transcriptions, slide film, and Super8 footage. The 4 hours and 50 minutes of recordings include elicited wordlists, phonetic tokens in carrier phrases, conversation, narratives, and songs. This collection includes recordings from some of the same speakers as are recorded in this collection and narratives that are similar to those in SUY1.

The materials collected by the Mother Tongue Centre Nepal will also be archived as an open access collection with PARADISEC (MTC1).[24] This collection includes 28 hours of original recordings, including songs, stories, and conversation from over 50 participants, with oral translations and careful respeaking for an hour of materials and written translations for over six hours. The transcriptions provide the possibility of drawing this collection together with the SUY1 collection to provide a broader analysis of narrative and song in Yolmo.

There are also a number of collections of closely related languages archived with PARADISEC. As mentioned in §6, Syuba is closely related to Yolmo. I have created a collection of Lamjung Yolmo materials (LG1).[25] This collection includes some materials that parallel the Syuba collection, including recordings of tasks such as Family Story, Hidden Objects, and 20 Questions (see §3.3.2). There are also recordings of the same tone elicitation sets (§3.3.1) and several earthquake narratives (§3.1.5.). Some parts of this collection are open access, and others can be accessed for research purposes. I have also published a sketch grammar of Lamjung Yolmo (Gawne 2016a) and a trilingual wordlist dictionary (Gawne 2011), as well as a PhD thesis that focuses on evidentiality, reported speech, and questions (Gawne 2013a). A collection of Anna Maria Hari's cassettes of Melamchi Yolmo from the 1980s has also been digitized and is available open access though PARADISEC (AH1).[26] The 20 cassettes resulted in 18.5 hours of audio. Hari (2010) published a grammar and a dictionary (Hari & Lama 2004) of Melamchi Yolmo. There is also a collection of 10 hours of open access audio recordings with BOLD respeaking and translations from Langtang, a variety that is closely related to Yolmo and previously undocumented (LAN1) (Slade 2014).[27]

3. The contents of the collection The materials in this collection can be broadly divided into different types, which will be discussed in turn. These include narratives

[23]This collection will be online in September 2017: http://catalog.paradisec.org.au/collections/MH1.
[24]http://catalog.paradisec.org.au/collections/MTC1
[25]http://catalog.paradisec.org.au/collections/LG1
[26]http://catalog.paradisec.org.au/collections/AH1
[27]http://catalog.paradisec.org.au/collections/LAN1

and conversations (§3.1), songs (§3.2), materials with a linguistic focus (§3.3), materials documenting the documentation (§3.4), and other materials (§3.5). I discuss the Ilam Yolmo materials in the collection specifically in (§3.6). I will mostly focus on the Syuba materials, and figures given will be specifically for the Syuba content, but also point out relevant Ilam Yolmo materials in each category.

The collection contains the following, in total, including Ilam materials:

- video recordings (114 MTS recordings, approx. 14.6 hours)
- audio recordings (214 WAV recordings, approx. 28 hours)
- ELAN annotation files (84 EAFs)
- FLEx files (3 XML files)
- images (535 JPGs)
- geolocation data (2 KML files)
- scans of notebooks (11 PDF documents)
- edited films (2 MOV files)
- picture books (2 PDF documents)
- academic papers (6 PDF documents)
- experiment data (2 bundles)
- metadata set (3 CSV files)
- administrative information (3 PDF documents)

The majority of recordings are open access, unless there were concerns about the privacy of the individual in relation to the content. This includes all notebooks and the metadata sheet for all participants, as personal information about individuals is included in these. One experiment data set is currently closed, and will be made open when the experiment results are published.[28] Geolocation data is also closed, as the data includes non-public paths between villages and households. The FLEx and ELAN files have been made open. These files are still being enriched, but since no documentation project is ever really complete, there is no easy-to-identify time where these files will be "ready" for open access. Therefore, these files come with an important caveat that there may be elements that are incomplete or unanalyzed. Users of these files should document download date and include it when citing the data (see §4.3).[29]

[28]There are also some recordings with Norpu Lama from 2010 and 2012. These were made before the consent model with a strong focus on open access was implemented, and as such remain on a more restrictive level of access than other recordings in the collection.

[29]They were originally left closed, however with feedback from reviewers, and discussion with colleagues, I have decided to make them open access. Conversations like these, and the opportunity to reflect on processes and expectations, are part of the benefit of producing a written summary of an online collection.

3.1 Narratives and conversation

3.1.1 Traditional narratives There are 14 recordings (84 minutes) of traditional narratives. These are mostly fables with animal characters, with the animals having personality types common to Nepali folktales. For example, in both the story of *Jackal and Pheasant* (SUY1-140126-15) and *Jackal and Horse* (SUY1-140128-03) (Figure 4), the jackal is a mischievous trickster. The characters in these stories are not always animals; for example, in *Misunderstood Children* (SUY1-140126-09), the story is about two children who speak in riddles and amuse the king with their cleverness. All of these narratives were recorded during the 2014 Firebird Foundation period. This was around the time that the MTCN documentation work was being done, and the stories recorded here were intended to supplement what had already been recorded in the other collection. Some of these stories have been translated into Nepali using oral translation techniques (§3.3.4), and many have accompanying ELAN transcriptions, interlinearizations, and translations. A subset of these stories have been included in the picture books published with NTU ADM students, which I discuss in §3.5.2. There were no traditional narratives recorded in Ilam.

Figure 4. Pasang Maya Lama telling the story of the Jackal and Horse SUY1-140128-03 07:47

3.1.2 Oral histories A number of participants, particularly senior members of the community, wanted to share narratives of Syuba history. There are 15 recordings (105 minutes) of these histories. Some touch on specific themes; for example, SUY1-160428-02 and SUY1-160428-03 are about the history of the local Buddhist *gompa* (temple) in Nobra, as told by Ringjin Lama. Some of these histories were recorded as conversations where older community members tell younger people about what life used to be like; in SUY1-160420-06, Lhamu tells Som Maya about the old days

(Figure 5), and in SUY1-160420-04, 75-year-old Larkel talks about his grandparents and great-grandparents with Sangbu Syuba.

Figure 5. Lhamu and Som Maya in SUY1-160420-06 11:11

There is also a recording of Ningmar Tamang of the Dongba clan discussing history with Dongba members in Lamjung (SUY1-160516-02) (Figure 6). Ningmar came to visit Lamjung with me in May 2016 and we recorded this conversation. In PARADISEC this recording is archived as LG1-160516-02 in the collection of Lamjung Yolmo materials.

Figure 6. Ningmar Tamang and Dongba clan men in Lamjung (SUY1-160516-02)

There is one oral history recording from Ilam, and it is also pertinent to the Syuba of Ramechhap. In SUY1-160505-02 Dorje Yolmo, a 95-year-old who lives in Tin Khutte, Ilam, talked about how his family moved from Ramechhap in the early 1930s when he was a young boy (Figure 7).

Figure 7. Dorje Yolmo talking about his youth (SUY1-160505-02 8:41)

3.1.3 Explicatory texts In this category I include both explicatory materials and footage of people performing these activities. There are six recordings (33 minutes) of people describing how people go about doing traditional activities, and nine recordings (58 minutes) of people performing various activities.

Explicatory texts include discussions about how to make alcohol and how to negotiate engagements for marriage (two topics that are traditionally closely related) (SUY1-141011-02), how to make butter and other dairy goods (SUY1-141011-04), and a description of the work of a local shaman (SUY1-160425-06) (Figure 8).

Videos of activities include practical domestic activities: bamboo mat weaving (SUY1-140129-04), butter churning (SUY1-160425-15 to 17), and harvesting honey from domestic hives (SUY1-160430-01). There are also two sets of videos recorded at social events in the villages; the first set are from a wedding (SUY1-160421-01) and the second set are from a community "Devi Puja" ceremony in Dhungare, performed at the start of the spring season (SUY1-160426-06). Many of these videos are accompanied by explicatory recordings where people describe the activities. The description of honey harvesting (SUY1-160430-05) and footage of opening the hive have been combined into a short edited documentary video (see §3.5.3).

There is one recording from Ilam where Pema Yolmo and Sange Yolmo discuss how weddings are celebrated in Ilam (SUY1-160505-19).

Figure 8. Norpu Tamang discussing the tools of a shaman (SUY1-160425-06 4:42)

3.1.4 Personal narratives Some participants shared narratives of their personal lives, which led to 11 recordings (56 minutes). All participants were made aware that the recordings were to be made openly accessible, and so these narratives represent people's public representations of their own lives, often touching on the hardships that they face.

There are also eight personal narratives from Ilam, including Mingmar, Tsenga Lamu, and Sange Yolmo discussing life in Ilam (SUY1-160505-10).

3.1.5 Earthquake narratives On April 25, 2015, a magnitude 7.8 earthquake struck Nepal. This earthquake, and the persistent aftershocks killed nearly 9,000 people in the country, and left thousands of others injured or homeless. There were very few serious casualties in Ramechhap, but 12 months after the initial earthquake, many were still living in temporary housing. We made 22 recordings (174 minutes) with 27 participants in which people shared their experience of the earthquake and its aftermath. These recordings were all made in 2016.

These recordings are an important documentation of a major event in local history. There are also other research groups that have been collecting earthquake narratives with other Tibeto-Burman language speakers in Nepal, including The Langtang Memory Project[30] with Langtang speakers, and The National Science Foundation (NSF) Grant for Rapid Response Research (RAPID) *Narrating Disaster: Calibrating Causality and Responses to the 2015 Earthquakes in Nepal*, which has made recordings with speakers of Tsum, Nubri, Lowa, Nar, Manange, Kuke, and Ghale.[31] The Syuba materials add to this collection of voices that are being heard about the earth-

[30] www.langtangmemoryproject.com

[31] e.g. Nubri: https://audio-video.shanti.virginia.edu/collection/nubri-reflections-2015-nepal-earthquakes

quakes, and can be included in cross-linguistic comparison of language use on a single topic.

There are no narratives about the earthquakes in Ilam Yolmo as the quakes themselves were only lightly felt in the area and after-effects were minimal.

3.1.6 Conversation There are 10 recordings (193 minutes) marked as conversations in the collection, although there are also other recordings with two or three speakers that have conversational data. Two recordings were made specifically where the camera was left on for an extended period during a morning in a house in one of the villages, to illustrate the kind of interactional rhythms in daily discourse (SUY1-160426-01, 13 minutes and SUY1-160426-02, 20 minutes) (Figure 9).

There are two specifically conversational recordings from Ilam, SUY1-160507-02 and SUY1-160507-03, both recorded with a group from the village of Namsaling, around two hours from where I was staying, who came to attend a funeral.

Figure 9. Morning conversation (SUY1-160426-01 17:56)

3.2 Songs There are 20 recordings of songs (around 90 minutes) in the collection. Songs are sung without musical accompaniment. These songs fall broadly into two main categories. The first type are songs that the singer has themselves written. The themes of these songs are a reflection of everyday life and concerns for the singer. Kabire sings of the hardship of life (SUY1-140127-04), Jit Bahadur sings of village life (SUY1-140127-05), and Pasang Maya's songs are about love of family (SUY1-140128-05). These songs are notable for their use of repetition, and epenthetic vowels to regulate rhythm.

The second type of song are traditional songs that people remember and sing in groups, such as those songs and dances that are often performed at weddings and other events. A group of Syuba women perform one such song (SUY1-140129-03)

while other members of the family perform acts of daily work, in a scene devised by the participants to demonstrate this type of dance (Figure 10).

Figure 10. Women singing a traditional song while dancing (SUY1-140129-03 2:17)

There are also songs in the Ilam collection. These are predominantly traditional songs from the Helambu area that they have learnt recently as they reconnect with other Yolmo groups. SUY1-160505-16 was performed by a group of women who started singing together in recent years.

3.3 Materials with a linguistic focus

3.3.1 Elicitation recordings There are 19 elicitation recordings (334 minutes) made at various points in the documentation process. These include a 100-word Swadesh list (SUY1-091125-01), as well as discussions of more specific lexical domains such as plants and animals (SUY1-160413-03). There are some elicitation sessions on syntax as well (SUY1-120318-01, SUY1-160413-01). These recordings do not include video.

There are several sets of words elicited for tone, both in isolation and in carrier sentences (e.g., SUY1-140128-07, SUY1-140204-01). These sets were used as the basis for one of the tone perception experiments discussed in §3.3.3. There are also comparable recordings with Lamjung Yolmo speakers in the Lamjung Yolmo collection (LG1-141106) that are being used to analyze and compare lexical tone in these two varieties.

There is also a 200-word Swadesh list recorded with Ilam Yolmo speakers (SUY1-160506-04) as well as a session where I discuss lexical items that are different between Ilam Yolmo and Syuba with Ningmar Tamang and Sonam Yolmo (LG1-141212-02).

3.3.2 Stimuli experiments There are 10 recordings (84 minutes) of people participating in stimulus-based activities. These are activities that I also ran with Lamjung

Yolmo speakers, as a way of establishing the distribution of evidential forms in controlled interactions (Gawne 2013a). One is the Family Story picture card story task, described in San Roque et al. (2012) (SUY1-160414-02) (Figure 11), and another two are tellings of the Jackal and Crow story described in Kelly & Gawne (2011) (SUY1-101011-01). There are also two rounds of the "Twenty Questions" game (SUY1-160415-01) and five recordings of the "Hidden Objects" stimuli tasks, both described in Gawne (2013a). There are no experiment recordings from Ilam.

Figure 11. Sangbu and Kali Syuba doing the family story picture task SUY1-160414-0219:00

3.3.3 Tone experiments Two different tone experiments were conducted during this project, to better understand the nature of the lexical tone in Syuba and its perception by speakers.

The first experiment was an auditory perception test as part of a collaborative project with Amos Teo. The tone recordings discussed in §3.3.1 were cut and presented as isolated tokens. Participants were then asked to perform a two-choice listening test, responding with which word they thought they were hearing. For example, *to* could either be 'rice' (*tó*) or 'stone' (*tò*) depending on the tone. 12 speakers listened to 120 tokens. The data are currently being analyzed. The bundle for this set is SUY1-experimenttone and includes the original stimuli materials and anonymized results. This bundle will be made open access when the experiment results are published.

The second experiment maps the perception of tone in relation to physical properties as part of a collaborative project with Suzy Styles at Nanyang Technological University. Participants listened to tone minimal pairs and matched them to two items of various parameters – for example a long rod and a short rod, a light and dark brown piece of cloth, or a heavy and light ball. Participants also matched the objects to a "kiki/bouba" pair, pairs of pure tone, and pairs of dog barks. This exper-

iment was done with 24 Syuba speakers and 24 native English speakers. The bundle for this set includes the auditory recordings, instructions for building the objects, as well as files for a 3D printable set, instructions for how to run the experiment, and anonymized participant results for all speakers. The bundle for this experiment is SUY1-experimentmetaphor and is also available through an Open Science Framework repository (Gawne & Styles 2017).[32]

Figure 12. Sabina Syuba using the Aikuma phone app to provide oral translation of a narrative (SUY1-images-5353).

3.4 Material documenting the documentation

3.4.1 Consent recordings For everyone who participated in the recordings, spoken consent discussions were also recorded and are included in the archive. These are mostly recorded in Nepali, although some include conversation in Syuba. Up until 2015 all recordings are made by me, but in 2016 Ningmar Tamang performed the consent interview, and I was present for all recordings. These recordings are all given the title "Consent", followed by the person's name.

These recordings were an opportunity for participants to also share their thoughts on the Syuba language, and the documentation project. For example, Sangita Tamang used the recording as an opportunity to discuss how her illness had prevented her

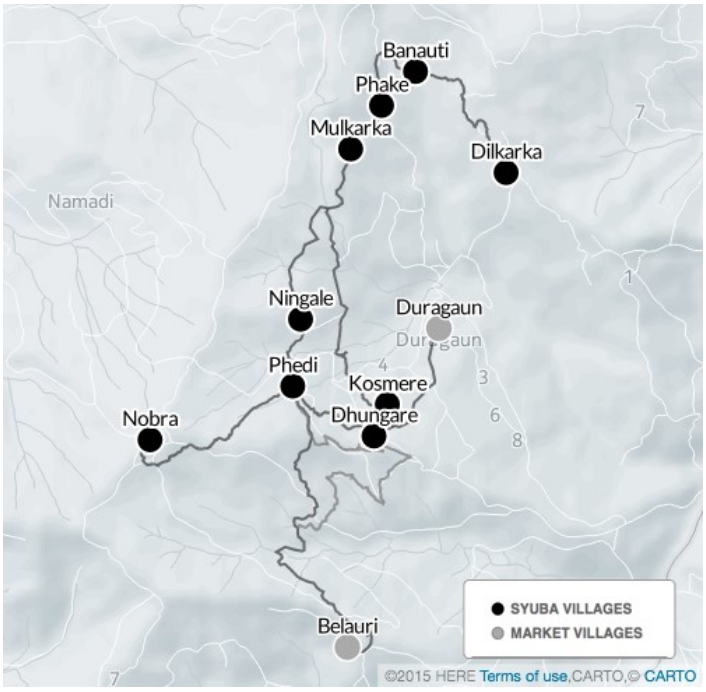
[32] http://osf.io/wt95v

from attending school as she wanted to (SUY1-160429-04) and Norpu Tamang used the recording to discuss how he was very happy the documentation was taking place, as he had very few tangible records of his own parents (SUY1-160425-05).

3.4.2 Images 535 photographs are archived with this collection. The majority of images illustrate daily life for Syuba speakers in Ramechhap and Kathmandu, and to document the project. The sequence number for the majority of photographs is the image number as it was saved onto my camera. As not all images taken were included in the archive, there are many breaks in the image number sequence.

3.4.3 Geolocation data There are two geolocation files in the collection, both in KML format, which is an XML file type for geolocation data. The first is a set of individual points for each of the villages that I visited during the documentation process. The second is a set of line data that marks the paths between the villages. This data was gathered on a Garmin ETrexH. The files were used to generate the map in Figure 13 using Carto,[33] an online mapping program. The geodata is available to other researchers, but as it includes geographic data on private local pathways, rather than local roads, it is closed access.

Figure 13. Villages in Ramechhap visited during the documentation project, and the paths between them.



[33] www.carto.com

3.4.4 Notebook scans There are two sets of notebooks, those that were made during elicitation and those that contain general observations about language and community life. These are in separate bundles in ELAR; the first is "Notes – language" and the second is "Notes – anthropological". These are bundled together in PARADISEC as SUY1-notebooks. The transcription notes have been used to build the ELAN transcripts. Both sets of notes also contain personal information about identifiable individuals and as such have been archived as closed items.

3.4.5 Other administrative documentation There are other administrative documents related to the project that are archived in the collection. The first set of these are the ELDP grant application (SUY1-administration-application) and the annual reports at the end of the first and second years of the project (SUY1-administration-annualreview1, SUY1-administration-annualreview2). These files have some personal information redacted. There is also the full CSV of the metadata of information about the collection (SUY1-metadata-01), a CSV with metadata about participants (SUY1-participants-01), and a CSV with metadata about the photographs (SUY1-images-001). The set about participants is archived as a closed file, as it contains personal information on members of the community.

3.5 Other activities and materials

3.5.1 Academic outputs Academic outputs that draw upon the Syuba documentation materials have been included in the archive. These include:

- An analysis of the lexical similarity between Melamchi Yolmo, Lamjung Yolmo, and Syuba (Gawne 2010) (SUY1-Publications-Gawne2010_lexicon)
- A more general discussion of the relationship between Yolmo and Syuba (Gawne 2013b) (SUY1-Publications-Gawne2013_YolmoKagate)
- A discussion of the digitization of the 1970s Syuba wordlist, and a comparison with the lexicography work on Lamjung Yolmo (Gawne 2014) (SUY1-Publications-Gawne2014_Lexicography)
- A discussion of the Aikuma app (Bird et al. 2014) (SUY1-Publications-BirdGawneEtAl2014_Aikuma)
- An analysis of the relationship between tone and intonation in Syuba and Lamjung Yolmo (Teo et al. 2015) (SUY1-publications-TeoGawneBaeseBerk2015_tone)
- A discussion of identity in relation to language and culture, as well as language and social group naming, across different Yolmo varieties and Syuba (Gawne 2016c) (SUY1-Publications-Gawne2016_identity)
- An overview of the historical and social context of the Syuba languages (Gawne 2017a)

- An analysis of the "kiki/bouba" test that was run with Syuba speakers (Styles & Gawne 2017)

This article will also be included in the collection, and future publications arising from work with the collection will likewise be added.

3.5.2 Story books Two storybooks were created by NTU students as part of the collaborative project with Joan Kelly in the School of Art, Design and Media. These have also been archived. These are a collection of eight stories and songs illustrated by Ng Xiao Yan (SUY1-picturebooks-01) and the story of the Jackal and Pheasant illustrated by Jolene Tan (SUY1-picturebooks-02).

3.5.3 Edited films There are two short documentary films made by Chouette Films with materials from this corpus. These were part of a larger project by the Endangered Language Archive (ELAR) to encourage engagement with endangered language documentation materials. The first (SUY1-documentaries-01) is a film about collecting honey from domestic bee hives, drawing on explicatory recordings from the corpus (§3.1.3). The second (SUY1-documentaries-02) weaves together edited clips from people's earthquake narratives (§3.1.5) with still images of the damage done in the villages. ELAN files are also included with the films with the text in Syuba (Devanagari orthography), Nepali, and English for subtitling.

3.6 Ilam Yolmo materials In early May 2016 I travelled with Ningmar Tamang and Sangbu Syuba to Ilam, on Nepal's eastern border with India, to meet with speakers of Ilam Yolmo. I had previously met Lapsang and Sonam Yolmo, two Ilam Yolmo speakers who visited Kathmandu for the Syuba dictionary workshop, in December 2014. This is the first documentation of Ilam Yolmo, which was first noted as a distinct variety in Thokar (2009).

There are a total of 36 recordings in the Ilam Yolmo set (6.3 hours). 13 of these recordings document oral consent. In eight recordings people narrate their personal histories. Dorje Yolmo's historical narrative about migrating from Ramechhap also includes many personal narrative details (SUY1-160505-02). There are also seven recordings of songs. These are mostly performed by groups of people, and in SUY1-160506-02 the participants dressed in Yolmo clothes and perform a dance while singing (Figure 14).

One recording involves the description of a traditional wedding ceremony (SUY1-160505-19), and three recordings are elicitations, mostly to provide lexical comparisons with Syuba and other varieties. There are also a number of conversation recordings.

Figure 14. Filming participants dancing and singing a traditional Yolmo song in costume in SUY1-160506-02 (SUY1-images-8605).

4. Next steps

4.1 Current work Current work focuses on furthering transcription of existing data, and analysis. Ningmar Tamang is continuing to transcribe and translate texts. These will be added to the archive incrementally.

My current academic work focuses on the use of co-speech gesture in the corpus (see Gawne 2016b; 2017b). The tone experiments are currently being analyzed and written up for publication as well (Styles & Gawne 2017 is the first). Work will soon commence on analysis of the evidential and other grammatical features. This is part of a larger project to understand the relationship between Syuba, Yolmo, and other closely related varieties, as part of a David Myers Research Fellowship at La Trobe that will conclude in 2020.

4.2 Future work and potential areas of collaboration A documentation archive of this scale offers a great many more topics of exploration than a single researcher can address in their own work. It is difficult to imagine just what uses others may find for a collection when it is made, but – as one reviewer for this article phrased it – it is dangerous to operate on an "if-we-build-it-they-will-come" assumption. In the

spirit of encouraging others to work with this data, I outline a number of research and non-research uses to which the corpus can be put.

First, there are language-internal features of Syuba that are still in need of even basic description. There is still much to be done describing grammar, discourse, and narrative in Syuba. Höhlig (1978) provided a taste of some of the features, but more research is needed. For anyone interested in undertaking a study of ergativity, reference tracking, or narrative structure in a Tibeto-Burman language, there is enough structured data here to begin an analysis. Another component of the collection which is currently neglected are the songs. The interaction of musical prosody and tone has not yet been analyzed, nor the relationship between rhythm and syllable structure. Many of these topics would suit research for a minor thesis.

Second, this collection has the potential to be used in cross-linguistic and typological investigations of linguistic phenomena. There are a number of elicitation and structured-task activities in this collection that replicate existing research methods, or are replicable by other researchers. The Family Story task (3.3.2) follows the same method as the 25 languages in the SCOPIC collection (Barth & Evans 2015). The earthquake narratives are similar to the NSF RAPID project discussed in §3.1.5 and can be used in cross-linguistic comparison with those seven languages. The elicited tone sets and other word-lists can also be used for cross-linguistic phonetic analysis.

Third, there are non-linguistic research questions that this corpus can address. The oral histories provide information that could be used in anthropological and historical research without necessarily focusing on linguistic structures. The earthquake narratives provide one larger collection of materials that can enrich a social history of natural disaster. The collection of folktales (§3.1.1) can be used to understand the cultural relationship between the Syuba, their Tibetan cultural origins, and contemporary life in Nepal.

Fourth, it is my hope that these materials can be of use in non-research contexts as well. §3.5.3 outlines how primary footage from earthquake narratives and honey collecting have been edited into short films for a general audience. All materials are archived with a license that allows for new materials to be created and shared (under the same Creative Commons provisions). Therefore, the video recordings can be incorporated into documentary films or other projects on a variety of topics. For example, in SUY1-141022-03 Sangbu talks about the effects of climate change on local agriculture, and in SUY1-160425-06 Norpu Tamang talks about local shamanism.

Finally, but perhaps most importantly, I hope that Syuba speakers will find ongoing uses for this collection. Some speakers have shared the videos from the collection on social media. Children have been using the illustrated books, and we plan to print more copies. Even if speakers do not access the materials in the immediate future, access to the internet is slowly reaching Nepal's remote villages. Ensuring the collection is digital and discoverable means that speakers and their families will be able to find these materials in the future, to use as they see fit without a need to contact me directly.

In writing this overview I hope that I have given an indication of the breadth of materials that are available to other researchers. I welcome anybody who is interested

in collaborating, or who has more specific questions about the collection, to contact me.

4.3 Quoting and using the data Users of any part of the collection should acknowledge Lauren Gawne as the principal investigator. Contributors who have collected, transcribed, or translated the data, or were involved in other substantial ways, should be acknowledged by name where possible. All information on contributors is available in the metadata. Where necessary, users should also acknowledge the Endangered Languages Documentation Programme, NTU Singapore, Firebird Foundation, The Awesome Foundation, and Stack Exchange as funders of the collection. The funding for specific items in the collection is indicated in the metadata.

Citation should be to the relevant level of specificity to allow others to be able to locate the point in the data the analysis is from.

- Reference to the collection in general: e.g. Gawne (2009a)
- Reference to a particular narrative: e.g. Gawne (2009a SUY1-140127-02) "Song: Village Life"
- Reference to a particular element of a narrative: e.g. use of reported speech marker *ló* in "Song: Village Life" (SUY1-140127-0 00:54) (Gawne 2009a)

Please cite the collection as: Lauren Gawne (collector), 2009; *Kagate (Nepal)* (SUY1), Digital collection managed by PARADISEC. [Open Access] [date of access] DOI: 10.4225/72/56E976A071650.

Citation of any analysis (FLEx and ELAN files) should include the date those files were accessed. The ELAR collection can be cited in addition, but the PARADISEC collection should always be considered the primary collection.

5. Summary This article has been written to provide an introductory overview of the Syuba collection of linguistic materials. There is a great deal of concern amongst linguists that the growing number of openly available archives will lead to predatory "data scooping". My greater concern is that we are building more and more language corpora that are languishing and under-utilized. The Syuba are eager for their language to reach more people. I am eager that these materials additionally be of use to researchers. In this article I have given an overview of the structure of the archive, and a "behind the scenes" look at the motivation, structure and use of the collection. In doing so I hope that it has made the collection more accessible.

Acknowledgements My enduring thanks go to the Syuba community for sharing the enthusiasm for their language with me, and collaborating on this corpus. Thanks especially to Sangbu Syuba and Ningmar Tamang for spending so much time working on the collection and transcription, and Norpu Lama for his early encouragement. Funding for the documentation of Syuba came from Stack Exchange, The Awesome

Foundation (Ottawa), The Firebird Foundation, NTU Singapore, and The Endangered Language Documentation Programme (ELDP). Many thanks to these organizations for their support and allowing us to develop this work over multiple grants. My collaborators on various parts of this project, including Amos Teo, Suzy Styles, and Joan Kelly added new perspectives and enriched the documentation process. Thanks to the staff at PARADISEC and ELAR for their support during the archiving process. Finally, thanks to Gary Holton and the particularly thoughtful reviewer for *LD&C*, who both shaped the final version of this paper.

References

Austin, Peter K. 2013. Language documentation and meta-documentation. In Jones, Mari C. & Sarah Ogilvie (eds.), *Keeping languages alive: Documentation, pedagogy and revitalization*, 3–15. Cambridge: Cambridge University Press.

Barth, Danielle & Nicholas Evans (eds.). 2017. Social cognition parallax interview corpus (SCOPIC): A cross-linguistic reference. *Language Documentation & Conservation*, Special Publication No. 12. Honolulu: University of Hawai'i Press. http://hdl.handle.net/10125/24739.

Bird, Steven, Isaac McAlister, Katie Gelbart & Lauren Gawne. 2014. Collecting bilingual audio in remote indigenous villages. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, 1015–1024. Dublin, Ireland: August 23–29, 2014.

Bouda, Peter & Johannes Helmbrecht. 2012. From corpus to grammar: How DOBES corpora can be exploited for descriptive linguistics. In Nordoff, Sebastian (ed.), *Electronic grammaticography*, 129–159. Honolulu: University of Hawai'i Press.

Gawne, Lauren. 2009a. *Kagate (Syuba), an endangered Tibeto-Burman language of Nepal*. London: SOAS, Endangered Languages Archive. http://elar.soas.ac.uk/deposit/0388.

Gawne, Lauren. 2009b. *Yolmo (also known as Helambu Sherpa, Nepal)* (LG1). Digital collection managed by PARADISEC. doi:10.4225/72/56E825B0B80EA.

Gawne, Lauren. 2010. Lamjung Yolmo: A dialect of Yolmo, also known as Helambu Sherpa. *Nepalese Linguistics* 25. 34–41.

Gawne, Lauren. 2011. *Lamjung Yolmo–Nepali–English Dictionary*. University of Melbourne, Australia: World Oral Literature Project.

Gawne, Lauren. 2013a. *Lamjung Yolmo copulas in use: Evidentiality, reported speech and questions*. Melbourne: University of Melbourne. Doctoral dissertation.

Gawne, Lauren. 2013b. Notes on the relationship between Yolmo and Kagate. *Himalayan Linguistics* 12(2). 1–27.

Gawne, Lauren. 2014. Similar languages, different dictionaries: A discussion of the Lamjung Yolmo and Kagate dictionary projects. In Zuckermann, Ghilad, Julia

Miller & Jasmin Morley (eds.), *Endangered words, signs of revival*, 1–11. Adelaide: AustraLex.

Gawne, Lauren. 2016a. *A sketch grammar of Lamjung Yolmo*. Canberra: Asia Pacific Linguistics.

Gawne, Lauren. 2016b. The interrogative palms-rotated gesture in Kagate, a Tibeto-Burman language of Nepal. Paper presented at The International Society for Gesture Studies (ISGS) Annual Conference 2016, Paris, July 18–22, 2016.

Gawne, Lauren. 2016c. My name is Maya Lama/Syuba/Hyolmo: Negotiating identity in Hyolmo diaspora communities. *European Bulletin of Himalayan Research* 47. 40–68.

Gawne, Lauren. 2017a. Syuba language context. *Language Documentation and Description* 13. 65–95.

Gawne, Lauren. 2017b. The 'nothing' gesture in Syuba: An example of the 'away' gesture family of conventional co-speech gestures in a Tibeto-Burman language of Nepal. Paper presented at iGesto, Porto, February 2–3, 2017.

Gawne, Lauren, Andrea L. Berez-Kroeker, Barbara F. Kelly & Tyler Heston. 2017. Putting practice into words: The state of data and methods transparency in grammatical description. *Language Documentation & Conservation* 11. 157–189. http://hdl.handle.net/10125/24731.

Gawne, Lauren & Suzy J. Styles. 2017. Cross-sensory perception for tone. *Open Science Framework*. doi:10.17605/OSF.IO/WT95V.

Grierson, George Abraham. 1909/1966. *Linguistic survey of India*. 2nd edn. Delhi: M. Banarsidass.

Hari, Anne-Marie. 1980. *Hyolmo songs, stories and grammar drills* (AH1). Digital collection managed by PARADISEC. doi:10.4225/72/56E9795C3C78B.

Hari, Anna Maria. 2010. *Yohlmo sketch grammar*. Kathmandu: Ekta books.

Hari, Anna Maria & Chhegu Lama. 2004. *Hyolmo-Nepālī-Aṅgrejī śabdakośa* [Yohlmo-Nepali-English dictionary]. Kathmandu: Central Dept. of Linguistics, Tribhuvan University.

Hedlin, Matthew. 2011. *An investigation of the relationship between the Kyirong, Yòlmo, and Standard Spoken Tibetan speech varieties*. Chiang Mai, Thailand: Payap University. Master's dissertation.

Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36. 161–195.

Höhlig, Monika (collector). 1973. Hoehlig Syuba recordings (MH1). Digital collection managed by PARADISEC. (Open Access).

Höhlig, Monika. 1978. Speaker orientation in Syuwa (Kagate). In Grimes, Joseph E. (ed.), *Papers on discourse* 50, 19–24. Kathmandu: Summer Institute of Linguistics.

Höhlig, Monika & Anna Maria Hari. 1976. *Kagate phonemic summary*. Kathmandu: Summer Institute of Linguistics Institute of Nepal and Asian Studies.

Kelly, Barbara & Lauren Gawne. 2011. *Don't forget the kids! Recording children's talk in language documentation*. Paper presented at the 2nd International Conference on Language Documentation and Conservation, Honolulu, Hawai'i, February 11–13, 2011.

Lewis, M. Paul & Gary F. Simons. 2010. Assessing endangerment: Expanding Fishman's GIDS. *Revue roumaine de linguistique* 55(2). 103–120.

Mother Tongue Centre Nepal (collector). 2014. Syuba recordings (MTC1). Digital collection managed by PARADISEC. (Open Access).

Mitchell, Jessica R. & Stephanie R. Eichentopf. 2013. *Sociolinguistic survey of Kagate: Language vitality and community desires*. Kathmandu: Central Department of Linguistics Tribhuvan University, Nepal & SIL International.

Reiman, D. Will. 2010. Basic oral language documentation. *Language Documentation & Conservation* 4. 254–268. http://hdl.handle.net/10125/4479.

Salffner, Sophie. 2015. A guide to the Ikaan language and culture documentation. *Language Documentation & Conservation* 9. 237–267. http://hdl.handle.net/10125/24639.

San Roque, Lila, Lauren Gawne, Darja Hoenigman, Julia Colleen Miller, Stef Spronck, Alan Rumsey, Alice Carroll & Nicholas Evans. 2012. Getting the story straight: Language fieldwork using a narrative problem-solving task. *Language Documentation & Conservation* 6. 135–174. http://hdl.handle.net/10125/4504.

Schembri, Adam, Jordan Fenlon, Ramas Rentelis, Sally Reynolds & Kearsy Cormier. 2013. Building the British Sign Language corpus. *Language Documentation & Conservation* 7. 136–154. http://hdl.handle.net/10125/4592.

Slade, Rebekah (collector). 2014. Langtang (Nepal) (LAN1), Digital collection managed by PARADISEC. [Open Access] doi:10.4225/72/574B120949E4C

SIL International. 2017. FieldWorks Language Explorer (FLEx) [Computer software]. http://fieldworks.sil.org/flex.

SIL International & HIS Nepal (eds.). 2016. *Syuba-Nepali-English dictionary*. Kathmandu: SIL International and HIS Nepal.

Styles, Suzy J. & Lauren Gawne. 2017. When does maluma/takete fail? Two key failures and a meta-analysis suggest that phonology and phonotactics matter. *i-Perception* 8(4). 1–16. doi:10.1177/2041669517724807.

Teo, Amos, Lauren Gawne & Melissa Baese-Berk. 2015. Tone and intonation: A case study in two Tibetic languages. In The Scottish Consortium for ICPhS 2015 (ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow: University of Glasgow. Paper number 0893.

Thieberger, Nicholas. 2009. Steps toward a grammar embedded in data. In Epps, Patricia & Alexandre Arkhipov (eds.), *New challenges in typology: Transcending the borders and refining the distinctions*, 389–408. Berlin: Mouton de Gruyter.

Thieberger, Nicholas & Andrea L. Berez. 2012. Linguistic data management. In Thieberger, Nicholas (ed.), *The Oxford handbook of linguistic fieldwork*, 90–118. Oxford: Oxford University Press.

Thokar, Rajendra. 2009. *Linguistic fieldwork in Jhapa and Ilam districts*. Paper presented at The Linguistics Society of Nepal Annual Conference, Kathmandu, Nepal, November 26–27, 2009.

Tournadre, Nicholas. 2014. The Tibetic languages and their classification. In Owen-Smith, Tom & Nathan Hill (eds.), *Trans-Himalayan linguistics, historical and descriptive linguistics of the Himalayan area*, 105–130. Berlin: Mouton de Gruyter.

Withers, Peter. 2012. Metadata management with Arbil. In Arranz, V., D. Broeder, B. Gaiffe, M. Gavrilidou & M. Monachini (eds.), *Proceedings of the Workshop Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR at LREC 2012*, 72–75. European Language Resources Association (ELRA).

Wittenburg, P., H. Brugman, A. Russel, A. Klassmann & H. Sloetjes. 2006. ELAN: A professional framework for multimodality research. In *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation.*

Woodbury, Anthony C. 2003. Defining documentary linguistics. *Language Documentation and Description* 1. 35–51.

Woodbury, Anthony C. 2014. Archives and audiences: Toward making endangered language documentations people can read, use, understand, and admire. In Nathan, David & Peter K. Austin (eds.), *Language documentation and description*, 19–36. London: SOAS.

Lauren Gawne
l.gawne@latrobe.edu.au
orcid.org/0000-0003-4930-4673